Case Study 3 ("Probability and stochastic processes 1")

1.

Discrete random variables

Discrete random variables are variables that can through obtained by counting. So its possible values are countable. Discrete variables in the Airbnb data are minimum_nights and availability_365.

Continuous random variables

Continuous random variables are variables whosse values can be obtained through measuring. Continuous variable takes all values in a given interval of numbers. Continuous variables in our Airbnb data are id, host_id, latitude, longitude, price, number_of_reviews, reviews_per_month, calculated_host_listings_count.

2.

In [9]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
```

Load data from csv

In [10]:

```
sgData=pd.read_csv("listings.csv")
```

Preprocess data

Change not a number value to zero

In [11]:

```
sgDataToZero = sgData.select_dtypes(include=[np.number])
sgDataToZero = sgDataToZero.fillna(0)
print(sgDataToZero.isna().values.any())
```

```
False
```

Change not a number value to mean

In [12]:

```
sgDataToMean = sgData.select_dtypes(include=[np.number])
mean = sgDataToMean.mean()
sgDataToMean = sgDataToMean.fillna(mean)
print(sgDataToMean.isna().values.any())
```

```
False
```

3.

Calculate mean

In [13]:

```
for i in sgDataToZero:
    print(i, ": ", sgDataToZero[i].mean())
```

```
id :  23388624.629568737
host_id :  91144807.40533705
latitude :  1.314192464904513
longitude :  103.84878745794845
price :  169.33299607942328
minimum_nights :  17.510054382192994
number_of_reviews :  12.807385860629822
reviews_per_month :  0.6796319716706684
calculated_host_listings_count :  40.60768938914885
availability_365 :  208.72631845200456
```

In [14]:

```
for i in sgDataToMean:
    print(i, ": ", sgDataToMean[i].mean())
```

```
id :  23388624.629568737
host_id :  91144807.40533705
latitude :  1.314192464904513
longitude :  103.84878745794845
price :  169.33299607942328
minimum_nights :  17.510054382192994
number_of_reviews :  12.807385860629822
reviews_per_month :  1.0436686735289245
calculated_host_listings_count :  40.60768938914885
availability_365 :  208.72631845200456
```

Calculate variance

In [15]:

```
for i in sgDataToZero:
    print(i, ": ", sgDataToZero[i].var())
```

```
id :  103310190502777.36
host_id :  6709099893759819.0
latitude :  0.0009349800983125476
longitude :  0.0019074744053256827
price :  115727.60260831242
minimum_nights :  1771.9567354764006
number_of_reviews :  882.5501705877178
reviews_per_month :  1.3240641068855823
calculated_host_listings_count :  4242.601195609099
availability_365 :  21351.064476743264
```

In [16]:

```
for i in sgDataToMean:
    print(i, ": ", sgDataToMean[i].var())
```

```
id :  103310190502777.36
host_id :  6709099893759819.0
latitude :  0.0009349800983125476
longitude :  0.0019074744053256827
price :  115727.60260831242
minimum_nights :  1771.9567354764006
number_of_reviews :  882.5501705877178
reviews_per_month :  1.0766218313631537
calculated_host_listings_count :  4242.601195609099
availability_365 :  21351.064476743264
```

4.

Yes, there is a significant change between data that are replaced by zero and replaced by mean. When comparing both mean results and variance results from both data that are replaced by zero and mean, we can see that there is a difference in reviews_per_month column. This is the only column that has a difference since the other columns do not have not a number value that can be replaced by either zero or mean.